AD_____

Award Number: DAMD17-00-1-0448

TITLE: Cox Regression Model for Interval-Censored Data in Breast Cancer Follow-up Studies

PRINCIPAL INVESTIGATOR: George Y.C. Wong, Ph.D.

CONTRACTING ORGANIZATION: Strang Cancer Prevention Center
New York, New York 10021-4601

REPORT DATE: July 2003

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	July 2003	Annual (1 Jul 02-30 Jun 03)

4. TITLE AND SUBTITLE
Cox Regression Model for Interval-Censored Data in Breast Cancer Follow-up Studies

5. FUNDING NUMBERS
DAMD17-00-1-0448

6. AUTHOR(S)
George Y.C. Wong, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Strang Cancer Prevention Center
New York, New York 10021-4601

E-Mail: gwong@strang.org

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 Words)

The overall objective of this research proposal is semi-parametric inference of the Cox proportional hazards (PH) regression model for a survival function $\Pr(X > x|Z = z) = S(x|z) = [S_o(x)]^{e^{z\beta}}$, where X is a time-to-event variable, which is subject to interval censoring, Z represents the covariates, S_o is a baseline survival function, and β represents the regression coefficients. One objective of our research is to develop asymptotic inferences of the generalized maximum likelihood estimators (GMLE) of β and $S(\cdot|z)$. A critical limitation with GMLE under interval censoring is that it is computationally feasible only for a small data set. Thus the focus of another aspect of our research is the investigation of a simple alternative to GMLE obtained by a two-step estimation procedure involving data grouping. In the third year of our research, we have established asymptotic normality for GMLE of β under both discrete and continuous censoring distributions. We have also introduced a diagnostic plot for assessing PH assumption and proposed a chi-square test for it. The results will be useful to breast cancer researchers pursuing chemoprevention intervention trials involving surrogate endpoint biomarkers, and genetic epidemiologists conducting studies on familial aggregation of breast cancer and related cancers.

14. SUBJECT TERMS
Breast cancer, interval-censored data, Cox regression model, maximum likelihood, two-step estimation, asymptotic properties

15. NUMBER OF PAGES
12

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinion, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

_X_ Where copyrighted material is quoted, permission has been obtained to use such material.

_X_ Where material from documents designated for limited distribution is quoted, permission has be obtained to use the material.

_X_ Citations of commercial organizations and trade names in this report do not constitute an official Department of the Army endorsement or approval of the products or services of these organizations.

_N/A_ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals", prepared by the Committee on Care and Use of Laboratory Animals of the Institute of Laboratory Resources, National Research Council (NIH Publication No. 86-23, Revised 1985).

_N/A_ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

_N/A_ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institute of Health.

_N/A_ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

_N/A_ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

George Wang

7/30/03

principal Investigator

A. TABLE OF CONTENTS

B. INTRODUCTION

Interval-censored (IC) data are encountered in three areas of breast cancer research. The most common application is in clinical relapse follow-up studies in which the study endpoint is disease-free survival. When a patient relapses, it is usually known that the relapse takes place between two follow-up visits, and the exact time to relapse is unknown. In statistics, we say relapse time is interval censored. Interval censoring is also encountered in breast cancer registry studies in which information on family history of cancer is updated periodically. The Strang Breast Surveillance Program for women at increased risk for breast cancer, for instance, has enlisted over 800 women with complete pedigree information which is verified and updated continuously. Family history data such as age at diagnosis of a specific cancer, or a benign but risk-conferring condition, are obtained from each registrant at each update. Time to a cancer event, and definitely time to first detection of a benign condition, are at best known to fall in the time interval between the last update and age at diagnosis. A third but increasingly important area of application of interval censoring is in breast cancer chemoprevention experiments or prevention trials, which involve the observation of one or more surrogate endpoint biomarkers (SEB) over time. The scientific question of interest here is the estimation of time for the SEB to reach a target value, and time from cessation of intake of a chemopreventive agent to the loss of its protective effect. Unfortunately, the exact values of both these time variables are known only to lie in between two successive assay inspection times. In a breast cancer follow-up study, we will often encounter covariates (for instance, tumor size and nodal status in a relapse study, and baseline SEB value in a chemoprevention trial).

Let X denote a time-to-event variable with distribution $F(x) = Pr(X \leq x)$, or equivalently, survival function $S(x) = 1 - F(x)$. In interval censoring, X is not observed and is known only to lie in an observable interval (L, R). In our previous DOD funded grant, we have made fundamental contributions to both the theory of the generalized maximum likelihood (GML) estimation of S, and the computation in connection with the inference of GML estimator (GMLE) $\hat{S}$ of S. These contributions are restricted to the case of univariate interval-censored data without covariates.

The Cox proportional hazards regression model [1] specifies that covariates have a proportional effect on the hazard function of X. This model provides powerful means for fitting failure time observations to a distribution free model and for estimating the risk for failure associated with a vector of covariates. It is extensively used for right-censored data. Finkelstein [2] applied the Cox model to analysis of interval-censored data. However, she

did not establish asymptotic properties of the GMLE of the parameters in the model and the approach is limited to small sample sizes due to computational difficulty .

Our interest in IC data with covariates is driven by needs arising from two related areas of breast cancer research at Strang. First, our investigators in the Strang Cancer Genetics Program want to study various patterns of familial aggregation of breast, ovarian and other forms of cancer using family history data from the Strang Breast Surveillance Program. Studies of familial early onset of breast cancer, breast-ovarian and breast-prostate associations will lead to IC data with covariates; therefore, a proper statistical procedure together with a feasible software to deal with such data are very much needed. Second, we conducted a one-year chemoprevention trial of indole-3-carbinol (I3C) for breast cancer prevention. In this prevention trial we monitored the levels of two SEB's, a urinary estrogen metabolite ratio and a blood counterpart, both of which are subject to interval censoring. An earlier dose-ranging study of I3C conducted by Wong *et al* [3] has been published.

The overall aim of this research proposal is to develop statistical inference for interval-censored data with covariates that are encountered in breast cancer chemoprevention trials employing surrogate endpoint biomarkers, and in breast cancer registry follow-up studies of familial aggregation of breast and other forms of cancer. Asymptotic generalized maximum likelihood theory under the Cox regression model will be investigated and computer software package for maximum likelihood inference will be implemented.

C. BODY

C.1. Model Formulation and Likelihood Equations.

Let $Y_{K,1} < Y_{K,2} < \cdots < Y_{K,K}$ denote the follow-up times for a patient who has made K follow-up visits, in a longitudinal follow-up study. Since the number of visits for each patient may vary, K is a random positive integer. For convenience, define $Y_{K,0} = 0$ and $Y_{K,K+1} = \infty$. The time-to-event variable of interest, X, is not directly observed; instead, it is known to lie in between two successive censoring time points $(Y_{K,j}, Y_{K,j+1})$, where j =0, ..., K. Note that X is left censored if $j = 0$, strictly interval censored if $0 < j < K$, and right censored if $X > Y_{K,K}$. The observable interval-censored data corresponding to X is given by

$$(L, R) = (Y_{K,i}, Y_{K,i+1}) \text{ if } Y_{K,i} < X \leq Y_{K,i+1},\ i = 0, 1, ..., K. \tag{1}$$

In addition to (L, R), we also observe a $p \times 1$ covariate vector Z. We assume that K and the $Y_{k,j}$'s are independent of (X, Z).

The Cox regression model for the survival function at $X = x$ given $Z = z$ is represented by

$$S(x|z) = [S_o(x)]^{e^{z\beta}}, \tag{2}$$

where $z\beta$ is the dot product of Z and β, $S_o(x)$ is a baseline survival function and β is a p-dimensional regression coefficient vector.

Let $I_i = (L_i, R_i, z_i)$, $i = 1, ..., n$, be a random sample of size n interval-censored observations with covariates. In terms of the original observed intervals, the likelihood function of S and b is given by

$$\text{L} = \prod_{i=1}^{n}((S(L_i))^{e^{bz_i}} - (S(R_i))^{e^{bz_i}}), \tag{3}$$

where S is a survival function, and b is a $p \times 1$ dimensional vector. The GMLE of (S_o, β) is a value (S, b) that maximizes (3) over all survival functions S and all $b \in \mathcal{R}^p$.

Since S_o places all probability mass on the innermost intervals of the I_i's (see Peto [4] or Turnbull [5]), it is often computationally simpler to express L in terms of innermost intervals.

We say that an interval A is an innermost interval of the I_i's if A is a nonempty finite intersection of one or more of the I_i's such that either $I_i \cap A = \emptyset$ or $I_i \cap A = A$ for each i. Suppose there are a total of m distinct innermost intervals $A_i = (\xi_i, \eta_i]$, where $\eta_i \leq \xi_{i+1}$ and $m \leq n$. Then the likelihood function (3) is equivalently given by

$$\text{L} = \prod_{i=1}^{n}[(\sum_{k>l_i} s_k)^{e^{z_i b}} - (\sum_{k>r_i} s_k)^{e^{z_i b}}], \tag{4}$$

where $l_i = \sup\{j : \eta_j \leq L_i\}$, $r_i = \sup\{j : \eta_j \leq R_i\}$ and $s = (s_1, ..., s_m)$ denote the vector of the probability weights. The log likelihood of (s, b) is

$$\mathcal{L}(s, b) = \sum_{i=1}^{n} \ln[(\sum_{k>l_i} s_k)^{e^{z_i b}} - (\sum_{k>r_i} s_k)^{e^{z_i b}}]. \tag{5}$$

Note that $(\sum_{k>r_i} s_k)^{e^{z_i b}} = 1$ if $r_i = 0$ and $(\sum_{k>l_i} s_k)^{e^{z_i b}} = 0$ if $l_i = m$.

C.2. Generalized maximum likelihood estimation.

A GMLE of (s, β) is a value of (s, b) that maximizes the likelihood function (5). We could follow the Newton-Raphson (NR) algorithm taken by Finkelstein [2]. However, this

would involve the inverse of a matrix of order $(m+p-1) \times (m+p-1)$. Since m can be potentially large when n is large, the unmodified NR algorithm is not feasible for a large data set.

We advocate a computationally simple approach by first grouping the original data (L_i, R_i) and then applying a two-step iterative scheme to obtain the two-step estimators (TSE) of S_o and β based on the innermost intervals corresponding to the grouped intervals.

In the **first** year of our research, we have successfully implemented a computer software to calculate the TSE's of S_o and β. A manuscript on the two-step computation scheme, including simulation studies investigating sensitivity of estimated values of TSE to partition sizes, is ready for submission to a statistical journal ([7]).

In our **second** year of research, we have applied our two-step estimation procedure to the Cox regression analysis of a long-term prognostic follow-up study involving 375 women with unilateral T1-2N0, T1-2N1 and T3-4 breast cancer. All the patients were treated at Memorial Sloan Kettering Cancer Center and the follow-up are being conducted at Strang Cancer Prevention Center. The main objective of the study is to assess the prognostic significance of bone marrow micrometastasis (BMM) in predicting relapse. Standard clinical variables including nodal status and tumor diameter were included in the Cox model. Although we have not yet established asymptotic normality to validate the P values that were reported for the study, our two-step Cox regression analysis gave strong indication that BMM was not as predictive of relapse as previously expected (Osborne and Wong [6]). We shall return to the BMM analysis when we fully establish the asymptotic normality of the GMLE of the Cox regression parameters. In our **second** year of research, therefore, we have moved ahead of our statement of work by making a start for Task 8. Since the BMM relapse follow-up study provides a complete and final data set that optimally satisfies our need of an empirical example to illustrate our asymptotic GML procedure for Cox regression, we have chosen to focus on this data set instead of the examples mentioned in Task 8.

Also, in the **second** year of our research, we have established consistency of the GMLE of β and S_o (and hence $S(\cdot|z)$) under the following assumptions:

AS1: S_o is arbitrary and each of the censoring variables,Y_1, ..., Y_k takes on finitely many values.

AS2: S_o is arbitrary and each of the censoring variables,Y_1, ..., Y_k is continuous and some regularity conditions are imposed on either S_o or the joint distribution function G of K, Y_1, ..., Y_K.

Specifically, under AS1 and AS2

$$Pr\{\lim_{n\to\infty} \hat{\beta} = \beta\} = 1, \tag{6}$$

and

$$Pr\{\lim_{n\to\infty} \sup_{t\in H} |\hat{S}_o(t) - S_o(t)| = 0\} = 1, \tag{7}$$

where H denotes the support set of Y_1, ..., Y_K. Note that $\hat{S}_o(t)$ is guaranteed to be consistent for $t \in H$, and not elsewhere. However, the set H is not necessarily a time interval (for instance, H may be a collection of discrete points). In order for the consistency results to be more useful, we have established that if S_o is continuous, and the support of Y_1, ..., Y_K is dense in $[0, T]$ for some $T > 0$, then $\hat{S}_o(t)$ is consistent for all $t \in [0, T]$. The practical implication of the denseness requirement is that pointwise consistency of $\hat{S}_o(t)$ would hold only if all the subjects in a follow-up study must be followed at very frequent close intervals.

We have also established similar consistency results for the TSE, with an added assumption that the maximal length of the partition interval tends to 0 as n tends to ∞. These results are summarized in Wong and Yu [7].

Asymptotic normality is the most crucial aspect of our research because it is needed in making confidence statements and in performing hypothesis testing. In the **third** year of our research, we have investigated asymptotic normality under assumptions

AS3. S_o is arbitrary and satisfies a monotonicity condition, and each of $Y_{K,1}$, ..., $Y_{K,K}$ takes on finitely many values;

AS4. S_o is as in AS3, and each of $Y_{K,1}$, ..., $Y_{K,K}$ takes on countably many values;

AS5. S_o is as in AS3, each of $Y_{K,1}$, ..., $Y_{K,K}$ is continuous and some regularity conditions are imposed on either S_o or G.

Asymptotic normality of GMLE or TSE is straightforward to establish under the finite assumption AS3. As for AS4 and AS5, we have carried out extensive simulation studies to guide our research. The studies suggest that both GMLE and TSE of β and S_o are asymptotically normal under AS4. However, only GMLE and TSE of β can be asymptotically normal under AS5. We have just completed theoretical proofs to substantiate our numerical studies. A manuscript is being prepared to report our findings (see Yu and Wong [8]). Our simulation studies suggest that under AS5 asymptotic inference for GMLE and TSE of S_o, and hence $S(\cdot|z)$ will have to be accomplished via a bootstrap method. We shall defer this aspect of research to the fourth and final year of our DOD grant.

Cox regression is appropriate only if proportional hazards (PH) assumption is satisfied by the data. Under the PH assumption, the log-rank test is most powerful. At present, a statistically useful diagnostic plot for PH assumption is lacking. Moreover, a formal significant test is not available. In the **third** year of our research, we have provided statistical solutions to satisfy both these needs. We are preparing a manuscript to report this particular piece of research, which was not proposed in the original statement of work (see Wong and Yu [9]).

D. KEY RESEARCH ACCOMPLISHMENTS

- We have implemented a statistical algorithm for computing GMLE of the regression coefficients β and the baseline survival function S_o.
- We have implemented a statistical algorithm for computing TSE of β and S_o.
- Computer programs for both GMLE and TSE calculations have been made available to the public via the internet.
- We have proved consistency of GMLE and TSE of β and S_o under both discrete and continuous assumptions about the censoring distribution G.
- We have performed extensive simulation studies to investigate the asymptotic properties of GMLE and TSE of β and S_o. Our results have provided strong evidence that S_o is **NOT** asymptotic normal when G is continuous.
- We have derived the asymptotic normal means and covariance matrices of GMLE and TSE of β.
- When G is finite or countably infinite, we have derived the asymptotic means and covariance matrices of GMLE and TSE of S_o.
- We have proposed a diagnostic plot for checking proportional hazards assumption for Cox regression and constructed a chi-square test for assessing this assumption.
- We have begun asymptotic GML Cox regression analysis of a long-term breast cancer follow-up study assessing the prognostic significance of bone marrow micrometastasis in predicting relapse in a cohort of 375 women.

E. REPORTABLE OUTCOMES

- An oral presentation of an abstract at 2002 ASCO Meeting ([6]).
- An abstract published in 2002 ASCO proceedings ([6]).
- A manuscript on computation of GMLE and TSE of Cox regression parameters ([7]).
- A manuscript on consistency and asymptotic normality of GMLE and TSE ([8]).

- A manuscript on assessing the appropriateness of proportional hazards assumption for Cox regression ([9]).
- Computer programs for calculating GMLE and TSE made available for the public via the internet site *http://www.math.binghamton.edu/qyu/index.html.*

F. CONCLUSIONS

In the three years of our DOD grant, we have successfully accomplished most of our research objectives in developing asymptotic generalized maximum likelihood inference of Cox proportional hazards regression model. We have developed statistical algorithms that can efficiently compute GMLE and TSE of the regression coefficients β and the baseline survival function S_o for any reasonable sample size. We have proved consistency of GMLE and TSE of β and S_o under both discrete and continuous assumptions about the censoring distribution G. We have established asymptotic normality for GMLE and TSE of β for G unrestricted. When G is continuous, we have numerically demonstrated that GMLE and TSE of S_o are not asymptotically normal. We propose to complete Task 5(c) and Task 6(c) in the fourth and final year of our DOD grant by investigating a bootstrap method for the asymptotic interval estimates of S_o.

Cox regression is appropriate only if proportional hazards (PH) assumption is satisfied by the data. We have proposed a useful diagnostic plot for PH assumption and validated a chi-square test for it.

In the fourth and final year of research, we shall complete a computer software for asymptotic confidence intervals and hypothesis testing for GMLE and TSE of β and $S(\cdot|z)$ (Task 6(c)). We shall also complete the data analysis of the BMM prognostic study.

The results which we have established will be useful to breast cancer researchers pursuing chemoprevention intervention trials involving surrogate endpoints biomarkers, and genetic epidemiologists conducting studies on familial aggregation of breast cancer and related cancers.

G. REFERENCES

[1] Cox, D.R. (1972). Regression models and life tables. *J. Roy. Statist. Soc. B,* 34 187-220.

[2] Finkelstein, D.M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, 42 845-854.

[3] Wong, GY, Bradlow, HL, Sepkovic, D, Mehl, S, Mailman, J, and Osborne, MP (1997). A dose-ranging study of indole-3-carbinol for breast cancer prevention. *Journal of Cellular Biochemistry Supplement* 28/29 111-116.

[4] Peto, R. (1973). Experimental survival curves for interval-censored data. *Appl. Statist.* 22, 86-91.

[5] Turnbull, B. W. (1976). The empirical distribution function with arbitrary grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser.* B, 38, 290-295.

[6] Osborne, MP and Wong GYC. (2002). Breast cancer bone marrow micrometastases: a long-term prognostic study of systemic tumor cell burden on relapse. *Proceedings of American Society of Clinical Oncology*, 21, #228.

[7] Wong, G.Y.C. and Yu, Q.Q. (2003). Estimation under the Cox regression model with interval-censored data. (Under preparation).

[8] Yu, Q.Q and Wong, G.Y.C. (2003). Asymptotic properties of the GMLE under the Cox regression model with interval-censored data. (Under preparation).

[9] Wong, G.Y.C. and Yu, Q.Q. (2003). A Test for checking Cox's model with a dichotomous covariate. (Under preparation).